International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 5, Issue 03, March-2019

PRIVACY PRESERVING IN DATA MINING: A REVIEW

[1]Ms. Sandhya Khanna, [2]Ms. Renu Tiwari

[1]*Compucom Institute of Information Technology & Management, Jaipur*
[2]*Compucom Institute of Information Technology & Management, Jaipur*

Abstract— To advance a well information from available information, number of techniques and methods has been developed in the area of data mining so far. On the other hand privacy is one of the most important properties of information that any system should satisfy. The secrecy of the information must be maintained while sharing the information among different un-trusted parties. Thus privacy plays a major role and also an important issue in most of the data mining applications. Privacy is main concern in the present technological phase in the world. Information security has become a dangerous issue since the information sharing has a common need. So secrecy is suitable an progressively significant problem in multiple data mining uses in many fields such as medical research, intelligence agencies, hospital records maintenance etc.

Keywords— Privacy preserving, Data mining, Classification, Models, Techniques.

I. INTRODUCTION

Data mining is one of the core processes in Knowledge discovery of Databases. It is the process of extracting interesting (which may be previously unknown, nontrivial) information. It similarly relates complex data around particulars cooperating specific's exact to secrecy. We should create definite not only the complex information have been secret, then too the suggestion channel has been unnavigable.

To illustrate the importance of the privacy in the data mining context we provide two examples of real applications. The first application is the health care area. Here a hospital would like to release the patient's data for an external third party for research purpose, who is typically not a trusted party. However, insurance companies have a great interest in knowing the health record of the patients and their parent's in order to increase the insurance on that family. The second application is recollected form marketing and market analysis. In this example, a retailing company would like to know the pattern of customer's choice in order to improve the sales trends. While this data is not important for many people, it would be important for the companies competing in the same market. So it is more important to provide an image of the data for required task without revealing data to the third party. It is also observed that the level at which the privacy of data is preserved is a major concern.

There are huge numbers of databases that hold numerous confidential information such that people access those data correlating various information from various databases. For e.g., suppose that the hospital issues the table, which does not overtly show the names of patients. [1]

II. CLASSIFICATION OF PRIVACY PRESERVING TECHNIQUES

The aim of secrecy conserving data mining is to improve data mining procedures lacking growing the hazard of data which recycled to produce systems. The issue of secrecy conserving data mining has been widely considered with the data mining unrestricted in latest years. A number of actual procedures for secrecy conserving data mining planned former. Most procedures usage specific form of renovation on the novel information in serial to implement the secrecy conservation. The renovated dataset is completed accessible for mining and should happen secrecy wants lacking misplacing the profit of mining. We categorize them into the accompanying three classifications:

A. The Randomization Method

The randomization method usages the information alteration techniques in sequence to generate reserved demonstrations of the records. In maximum of the circumstances, the specific records cannot be improved, and then combined disseminations can be improved. These circulations can be used for data mining resolves.

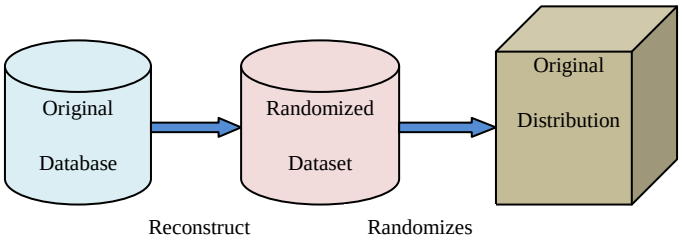

Figure 1 Model of Randomization.

Advantages:
It is easily implemented in data collection phase and all the other knowledge is not required and there is no need of server.
It is useful for hiding individual sensitive data.
Disadvantage:
One of the main disadvantages in Randomization is it treats all the records equally and reduces the utility of the data.

B. Anonymization
To defend particulars integrity while discharging complex data, data containers regularly eliminate open identifiers. These data may accidentally disclose the sensitive data about the individuals. The risk of linking the private information is handled by various privacy preserving techniques. The main concern is sensitive information which should not be disclosed. There are 2 categories of releases like, integrity discovery and attribute discovery. Identity disclosure happens when an individual is uniquely identified from the published data. Trait revelation happens when the data of an individual can be surmised from the published information. The number of secrecy models are conversed which are prospered in explaining the concerns like attribute discovery and integrity discovery using conserving personal information.

C. Distributed Privacy Preservation
Distributed procedures for privacy-conserving data mining is to convey out a combined calculation over the complete dataset deprived of conceding the privacy of the individual datasets within the different participants. The goal of safe multiparty calculation is to allow parties to convey out some disseminated computing efforts in a safe way. There are two usual sets for privacy-conserving data mining. In first setting, the information is circulated between at least two changed gatherings; the objective being to run an information mining calculation on the mix of the gatherings' databases using not agreeing any party to vision any other individual's personal information. In the 2nd setting, such arithmetical information i.e. to be free may hold intimate information. [2]

III. PRIVACY PRESERVING MODELS

In data sharing, various approaches are used to design the different privacy preserving models. In Privacy-preserving data mining (PPDM), which generally examined 4 types as ensues:

a. Trust Third Party Model
b. Semi-honest Model
c. Malicious Model
d. Further Models-Incentive Compatibility

a. Trust Third Party Model
The trustworthy 3rd party is executing the calculation and provides only the outcomes of the calculation. The goal of secure protocols is to reach this same level of privacy preservation, without considering the problem of finding a third party that assumes everyone is the faiths.

b. Semi-Honest Model
In the Semi-Honest Model, each party performances as per a semi-honest one. It ensues the instructions of the protocol with its exact i/p, then when the protocol is open to usage at all it realizes throughout the implementation of the protocol to cooperation the security.

c. Malicious Model
In the malicious model, restrictions are not placed on any of the participants. So uncertainty one party is finally able to spoil in at all movements it contents. All in all, these models are very hard to create effective conventions that are as yet legitimate under the malignant model.

d. Other Models - Incentive Compatibility
In the cryptographic unhindered, the semi-genuine and the mean models have been great inspected. 1 instance is the exciting privacy preserving concept is reason compatibility. A reason companionable protocol revenues that a unprincipled party is whichever held or other grieves an financial harm.

The above four privacy preserving models are considered the two basic privacy preserving models such as real model and ideal model. [3]

V. LITERATURE SURVEY

Abdullah H. Alsharif et.al. [2016] Data mining can be definite by way of the use of difficult apparatuses of data enquiry to determine formerly indefinite associations and designs in huge datasets. The apparatuses may comprise mathematical algorithms, arithmetical models and machine knowledge procedures. Therefore, data mining comprises techniques that enable more processes than data collection and management, including data analysis and prediction. Healthcare databases have huge amounts of data, and with effective analysis tools, a great deal of hidden knowledge may be discovered.

Therefore, data mining can be particularly useful for analyzing and extracting hidden knowledge in huge amounts of data such as those obtained from smokers. Data mining has found an application in the healthcare system. [4]

Bhoomika R Mistry et.al. [2015] have planned algorithm is the allowance of MDSRRC algorithm, which covers several Right Hand Side objects. In Planned effort MDSRRC algorithm workings on the disseminated database. They have presented investigational outcomes in links by MDSRRC algorithm in particular database and MDSRRC algorithm in disseminated database. They have tried to extend the MDSRRC algorithm in distributed database and also try to the MDSRRC algorithm, increase the efficiency and reduced the side effect by minimizing the modification on distributed database. [5]

Elham Hatefi et.al. [2014] have planned a novel system to cover complex quantifiable suggestion instructions which is built on rounded optimization method. In this technique, there are fuzzy suggestion instruction covering used for expressed as a rounded optimization problem and tests have been approved out on the actual database. The results of experiments indicate that the proposed method performed well exiting methods at this field in the term of percentage of missing rules and it changes made in the dataset of database. The results showed that all sensitive rules were made hidden with the proposed method. [6]

Nivetha .P.R et.al. [2014] Data mining is a technique where massive amounts of both sensitive and non-sensitive data are collected and examined. While distributing such private data, privacy preserving becomes an important issue. Several techniques and methods have been presented in privacy preserving data mining to accept this problematic. The main intention of privacy preserving is to extract the knowledge without disclosing private data and it also concerns about the sequential release of data. Sequential data helps in predicting the next occurrence which leads to violating the privacy of individual data. In this paper, we briefly surveyed sequential pattern hiding, k-anonymity, and data perturbation and secure sum computation techniques to address the issues of privacy preserving data sharing. [7]

S Kasthuri et.al. [2013] have proposed a new method to detect the sensitive items to hide sensitive association rules. This planned process treasures the connected or regular item groups and produces the suggestion instructions. It used the concept of representative association rules to detect sensitive items. In the proposed method, representative association rules are computed from association rules. [8]

Komal shah et.al. [2012] have arranged 2 calculations, ADSRRC and RRLR for covering such complex suggestion instructions. Equally algorithms are prepared for overwhelmed restrictions of extra instruction covering algorithm DSRRC (Decrease Support of Right Hand Side objects of Rule Cluster). ADSRRC algorithm overwhelmed restriction of several arranging in database and it chooses operation to be altered built on altered conditions than DSRRC algorithm.[9]

Ashish Bindra et.al. [2012] Publicists and large data mining specialists similar to are today are distributing by composite datasets of collective variation (first and third party data), size (actions, impersonations, connects), and rapidity (real time bidding). Generating analytical models to modify supporter desires and movement analytics to display besieged ads to operators who are maximum probable to change has convert progressively exciting. Promoters often set users in a section definite using a given group of demographic or communicative attributes. Some sections are often very bare. "Look-Alike Modeling" allows promoters to improve the goal section by using predictive models to increase the section membership by conveying a possibility score to users that did not explicitly belong to that segment based on the original segment definition. [10]

Jian Wang et.al. [2009] Privacy protection becomes a significant problem in the progress technique of data mining process. Privacy Protection Data Mining has developed progressively general since it agrees privacy-sensitive data distribution for enquiry resolves. That's why people have developed progressively disinclined to share their information; as a result regularly the person rejects to share their information or delivers the incorrect data. In turn, some issues in the information gathering can disturb the achievement of data mining, which based on the actual data in sequence to produce expressive outcomes. In current years, the prevalent obtain ability of private information has composed data mining an important issue of privacy. Various modes have been planned to reserve the data mining of a multidimensional data record currently. This letter obviously means to replication several concealment that shields data mining machineries and before changes onward to examine the assets and inadequacies of these technology. [11]

IV. PRIVACY PRESERVING TECHNIQUES

Use of private information for any movement lacking any data to the apprehension is word as Public Apprehension. In order to understand this thing consider a user never want to share their personal information with other without his permission, but it is done by the information holding organization, then this is called as public concern. Mostly it divided into few categories such as bank, work place, legal organizations; health care departments are most trustful for customer information privacy. While in case of credit-card companies, some of social site they are least trusted companies. PP procedures can be ordered dependent on the insurance strategies utilized by them, as appeared in Fig. 2.

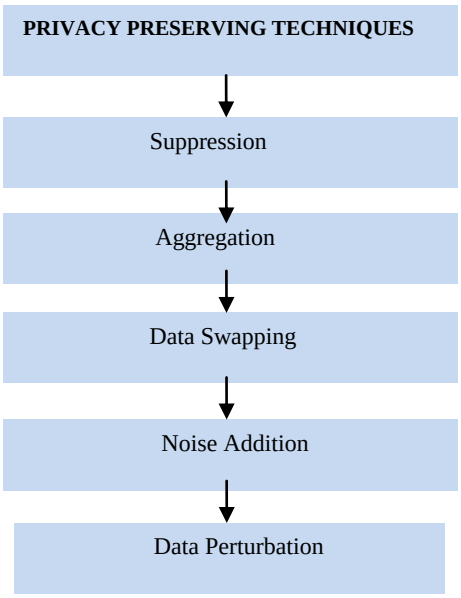

Fig. 2: Dissimilar Privacy Preserving methods.

A. Data Perturbation

Data is directly modified in this technique so it comes under data modification category. It is a classification of data transforming approaches that provide privacy for the discriminating or sensitive data from data miners. Here, some patterns or class of data is considered as the sensitive information which get modify by the different techniques of privacy preserving. So released data is containing inaccurate data where sensitive information is modified. While doing change it is required to do bother in the data having same measurements, as various qualities get straightforwardly go about as exceptions.

Advantage: One unused of the information make through disconcertion in the dataset in this way.

Disadvantage: A selection of same value for replacement can be easily identified by the miner. But replacing different value tends to lead the selection difficult.

B. Noise Addition

This method is put on numeric information only as per noise may be create using certain noise creating function for example Gaussian function. Here data quality is maintained by the technique so it looks like original, while privacy of the information is maintained. Noise addition also term as Randomization. The randomization method defends owner's information via letting randomly disturb their records formerly sending them to server, when pleasing gone actual data and comprising such noise. The original circulations of a disturbed information set can be irregular in case that the circulations of the equivalent novel data set and/or the circulations of the further noise is not multi-variant usual. So replies to queries connecting percentiles, sums, provisional means, etc. such noise sums methods are problematic Disquiet method, Random Disquiet Method and All Leaves Probabilistic Disquiet Method.

Advantage: Addition of noise is simple as based on some kind random function such as Gaussian.

Disadvantage: This can only perturb the numeric attributes while textual data remain same.

C. Data Swapping

Data swapping methods mostly request for process that retains all novel values in the data set, at the similar time the record re-identification is very tough. Data swapping revenues swapping the novel information set via any other one. Here a few unique qualities having a place with a touchy characteristic are traded between them. This swapping may be complete hence therefore that the t-order indicators of the novel data set are conserved. A t-order fact is a figure that can be created after precisely t attributes. A new concept called approximate data swap was introduced for practical data swapping. It calculates the t-order occurrence table after the novel data set, and discovers out an original data set through almost the equal t-order occurrence.

Advantage: Here, originality of the dataset is maximized. This is better approach to increase the frequency of the non-frequent patterns.

Disadvantage: Generation of patterns is time taken then to order is needs to be accurate to hide sensitive data.

D. Aggregation

Generalization of the available data is also term as aggregation. This provides privacy of individual information by replacing a group of information with a single before releasing it. In other words aggregation replaces k number of session by its representative session. Some attribute value in the dataset is consequent via pleasing the usual of the group of similar attributes data. At this time one apprehension will stand up that unused of k-number of novel records through an aggregated any creates data loss. In sequence to decrease this loss of data clustering of the records is compulsory, where extent of cluster change to condensed for decreasing the data damage. But by responsibility this trespasser can

estimation the balance of data loss in the data. Complete risk of the revealed data is remnants same. Another technique of combination or simplification is conversion of quality standards.

Advantage: It impeccably leathers the detail data of a discrete and producing information after the mutual data is not conceivable.

Disadvantage: Aggregation is done here but it tends to make an information loss. This loss of information makes that dataset fruitless.

E. Suppression

Complex data values are removed or repressed former to the relief of a information in suppression method. In this method sensitive information is removed from the dataset which is identified by the system. A Sensitive value is predicted by an intruder through various approaches. For example, some inbuilt identifier can identify the sensitive information which needs to be perturbed. Therefore, sufficient number of attribute values should be suppressed in order to protect privacy. However, suppression of attribute values results in information loss. A significant issue in suppression is to minimize the information loss by minimizing the number of values suppressed. For some dataset like patient medical diagnosis suppression is done by adding noise so chance getting accurate data is very less.

Advantage: Now, complex data is totally secreted after the dataset. Intruder has no casual for any type of mining.

Disadvantage: In this method removing of attribute leads to high information loss and because of this weakness suppression method is rarely use for perturbation. [12]

VI. CONCLUSION

In today's world, privacy is the major concern to protect the sensitive data. At present, privacy preserving in secure data sharing is at the stage of development. People are very much concerned about their sensitive information which they don't want to share. It review discloses that the secrecy of the information can be conserved at multiple levels. The methods and techniques which we are applying must be an efficient one that will not result in high information loss.

REFERENCES

[1] A.S.Shanthi and Dr. M. Karthikeyan, "A Review on Privacy Preserving Data Mining", 978-1-4673-1344-5/12/$31.00 © IEEE 2012.

[2] K.Saranya, K.Premalatha and S.S.Rajasekar, "A Survey on Privacy Preserving Data Mining" IEEE sponsored 2nd international conference on electronics and communication system (ICECS 2015).

[3] Ms. Dhanalakshmi.M and Mrs.Siva Sankari.E, "Privacy Preserving Data Mining Techniques-Survey", ICICES - S.A.Engineering College, Chennai, Tamil Nadu, India 2014.

[4] Abdullah H. Alsharif and Nada Philip, "Data mining technique for the Enhanced Smoking Cessation management system (Smoke Mind)", 978-1-5090-5579-1/16/$31.00 © IEEE 2016.

[5] Bhoomika R Mistry & Amish desai, "Privacy preserving heuristic approach For Association Rule Mining in Distributed Database" IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication Systems ICIIECS 2015.

[6] Elham Hatefi, Abdolreza Mirzaei & Mehran Safayani, "Privacy Preserving Quantitative Association Rule Mining Using Convex Optimization Technique", 7th International Symposium on Telecommunications (IST'2014).

[7] Nivetha .P.R and Thamarai Selvi .K, "Techniques for Privacy Preserving Data Sharing: A Survey", Department of Computer Science & Engineering Dr.N.G.P Institute of Technology Coimbatore, India 2014.

[8] S Kasthuri and T Meyyappan, "Detection of Complex objects in Market Basket Database with Suggestion instruction Mining for Secrecy Conserving" Records of the 2013 International Session on design Appreciation, Informatics and Mobile Engineering, February 21-22, 2013.

[9] Komal shah, Amit thakkar & Amit ganatra, "Suggestion Instruction Covering with Heuristic Method to Decrease Lateral Properties & cover several R.H.S. Items" International Journal of Computer uses (0975 – 8887) Volume 45– No.1, May 2012.

[10] Ashish Bindra, Sreenivasulu Pokuri and Ankur Teredesai, "Disseminated Vast Promoter Data Mining", IEEE 12th International Discussion on Data Mining Workshops 2012.

[11] Jian Wang, Yongcheng Luo, Yan Zhao and Jiajin Le, "A Review on Secrecy Conserving Data Mining", 1st International Workshop on Database Technology and uses 2009.

[12] Bhavna Vishwakarma, Huma Gupta and Manish Manoria, "A Survey on Privacy Preserving Mining Implementing Techniques", Symposium on Colossal Data Analysis and Networking (CDAN) 2016.